4-2012

Some drivers of test item difficulty in mathematics : an analysis of the competency rubric

Ross Turner
ACER, ross.turner@acer.edu.au

Ray J. Adams
ACER and University of Melbourne

Some Drivers of Test Item Difficulty in Mathematics – an Analysis of the Competency Rubric

Ross Turner

Australian Council for Education Research

and

Raymond J Adams

Australian Council for Education Research & University of Melbourne

Introduction

This paper is concerned with the empirical validation of the competency rubric described in the previous paper. Using items developed for the PISA 2012 survey, and data collected as part of an extensive field trial of the PISA tasks conducted during 2011 in some 67 countries, we use multidimensional Rasch modelling and latent regression to examine the following three questions:

1. What is the level of agreement among raters when they apply the competency rubric?
2. Does each of the competencies capture different dimensions of cognitive complexity in the tasks?
3. To what extent do ratings of the cognitive complexity account for (predict) the difficulty of the tasks for students?

For this scale evaluation and validation study we are using a pool of 196 mathematics tasks that were trialled in 2011 for use in PISA 2012. Each of the tasks was rated according to its demand for activation of the six competencies, by five mathematics educators who were intimately familiar both with the rubric and with the tasks. Four of the raters were members of the PISA 2012 Mathematics Expert Group, the group responsible for guiding the development of the PISA test, and the fifth member was leader of the PISA mathematics test development team.

Information concerning the difficulty of the tasks for 15-year-old students was obtained from the trial of the tasks in Australia in 2012. The trial was designed so that each of the 196 items was administered to approximately 50 randomly sampled students. While each student was only required to respond to about 50 tasks, a matrix sampling approach was used to ensure the tasks were all linked for the purpose of item response theory scaling.

To address the research questions we adopt a Rasch-based (Rasch, 1960) item response modelling approach in which we view each tasks as being characterised according to its six competency demands and its difficulty. These seven characteristics are seen as distinct, but likely correlated, latent dimensions.

More formally we assume that each task, *i*, can be characterised by a vector of seven latent characteristics, $\Theta_i = (\theta_{i1}, \theta_{i2}, \theta_3, \theta_{i4}, \theta_{i5}, \theta_{6i}, \theta_{i7})^T$ and we use a multidimensional rasch partial credit model to estimate the covariance structure of these latent variables. The model we employ is the multidimensional random rasch model of Adams, Wilson and Wang (1997) and we estimate it using the software ConQuest 3.0 (Adams, Wilson and Wu, 2012). Under this model we will assume that the ratings of each of five raters are indicators of the competency characteristics of the tasks and the responses of students to the tasks will be used as indicators of the task difficulty.

Results

Table 1 gives the frequencies of the rating provided by each of the raters for each of the competencies. In each case the total number of ratings is 196, the number of tasks that are being used in all analyses reported here. The total is the sum of the rating across raters in each category divided by five and then rounded. This metric was chosen to aid comparison with the distribution of ratings by individual raters. If the category labels (0, 1, 2 and 3) are used as *scores* of the

competency demands of the tasks the mean that is reported is the average score across all tasks provided by each rater for each competency.

Table 1: Frequencies for the Ratings by each Rater and for each Competency

Competency		*Raters*					
		CB	**KS**	**MN**	**RT**	**WB**	**Total**[1]
Communication	**0**	40	19	59	23	23	33
	1	103	136	86	99	125	110
	2	42	40	48	58	46	47
	3	11	1	3	16	2	7
	Mean[2]	1.12	1.12	0.97	1.34	1.14	1.14
Devising strategies	**0**	46	99	65	34	49	59
	1	96	64	91	39	107	79
	2	49	29	38	97	39	50
	3	5	4	2	26	1	8
	Mean	1.07	0.68	0.88	1.59	0.96	1.04
Mathematising	**0**	74	97	61	49	79	72
	1	73	70	105	74	103	85
	2	40	24	26	57	13	32
	3	9	5	4	16	1	7
	Mean	0.92	0.68	0.86	1.20	0.67	0.87
Representation	**0**	49	80	78	36	40	57
	1	80	71	98	81	128	92
	2	60	38	18	67	26	42
	3	7	7	2	12	2	6
	Mean	1.13	0.86	0.71	1.28	0.95	0.99
Symbols and formalism	**0**	52	74	80	26	95	65
	1	87	79	79	87	77	82
	2	50	33	33	67	22	41
	3	7	10	4	16	2	8
	Mean	1.06	0.89	0.80	1.37	0.65	0.96
Reasoning and argument	**0**	40	76	69	36	66	57
	1	90	62	105	76	106	88
	2	55	46	17	66	23	41
	3	11	12	5	18	1	9
	Mean	1.19	0.97	0.79	1.34	0.79	1.01

Notes:

1. Total is the sum across raters of the number of ratings in this category divided by five and then rounded to the nearest whole number.
2. Mean is average *score* of the ratings, that is the sum of the category level (0, 1, 2 or 3) times the count for each category divided by the total number of ratings

Typically rater RT is the most lenient, followed by CB. RT is about ½ of a category higher on average. MN is the harshest on average, but close to the remaining two. Figure 1 shows this in graphical form.

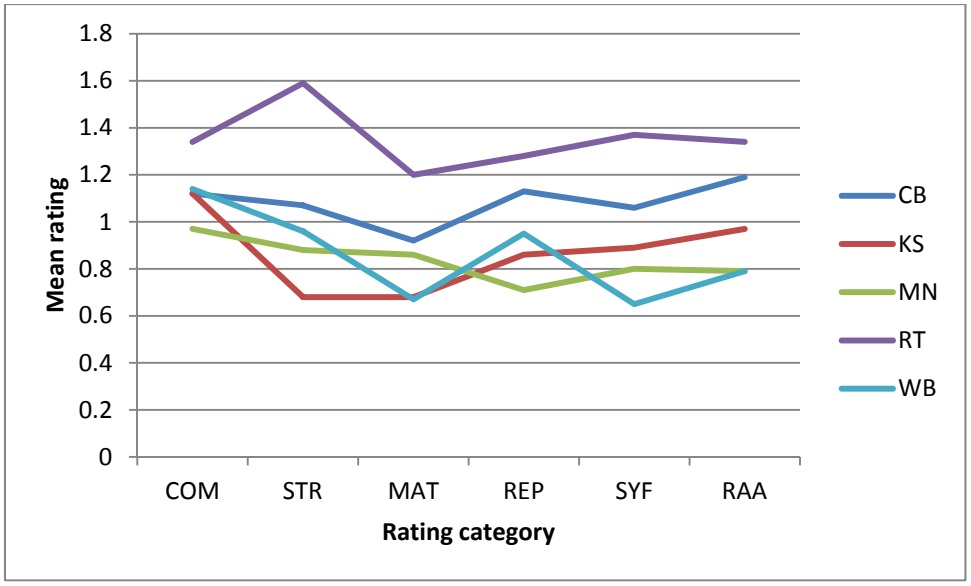


Figure 1: Mean Rating by Rating Category for each Rater

On average for these items, *Communication* attracts the highest rating for four of the raters, and *Mathematising* the lowest. At face value, this might indicate that many of the tasks considered involve some communication demand, whereas fewer tasks involve mathematising. This could be a feature of the tasks chosen, or of the scheme's application by these raters.

The frequency with which these raters use the various available levels varies quite a bit across categories and among raters, but it is clear that most of the raters used the highest rating level relatively rarely, certainly at a much lower rate than the other available rating levels.

Agreement Among Raters

Table 2 shows two indices of pairwise agreement between the raters for each of the competencies. In the tables the values given above the diagonals are the percentages of tasks that are given identical ratings by the two raters. For example raters CB and KS rated 52% of the tasks identically on the *Communication* competency. The values given below the diagonals are the product moment correlations between the sets of ratings provided by pairs for raters. For example the correlation between the *Communication* competency ratings of CB and KS is 0.35.

Table 2: Correlation and Percent Exact Match Indices of Agreement between Raters

Communication

	CB	KS	MN	RT	WB
CB		52.0%	52.6%	60.7%	55.6%
KS	0.35		44.4%	48.5%	52.6%
MN	0.53	0.40		45.4%	50.0%
RT	0.64	0.28	0.55		57.7%
WB	0.45	0.18	0.52	0.48	

Devising strategies

	CB	KS	MN	RT	WB
CB		46.9%	52.0%	44.4%	50.5%
KS	0.46		51.5%	28.6%	46.9%
MN	0.47	0.46		32.1%	52.0%
RT	0.57	0.42	0.44		33.2%
WB	0.37	0.31	0.49	0.34	

Mathematising

	CB	KS	MN	RT	WB
CB		45.9%	57.7%	49.5%	45.4%
KS	0.32		37.8%	45.4%	51.5%
MN	0.42	0.33		58.2%	49.0%
RT	0.66	0.36	0.59		43.9%
WB	0.33	0.10	0.39	0.44	

Representation

	CB	KS	MN	RT	WB
CB		43.9%	45.9%	66.8%	54.1%
KS	0.22		59.7%	50.0%	52.6%
MN	0.22	0.57		50.0%	61.7%
RT	0.56	0.41	0.54		55.1%
WB	0.50	0.53	0.57	0.56	

Symbols and formalism

	CB	KS	MN	RT	WB
CB		54.6%	50.0%	56.6%	52.6%
KS	0.57		55.6%	46.4%	61.7%
MN	0.49	0.55		37.8%	58.7%
RT	0.68	0.59	0.60		39.8%
WB	0.56	0.63	0.58	0.56	

Reasoning and argument

	CB	KS	MN	RT	WB
CB		43.4%	48.0%	73.0%	50.5%
KS	0.41		47.4%	40.3%	47.4%
MN	0.46	0.51		39.8%	55.6%
RT	0.75	0.42	0.53		41.8%
WB	0.52	0.37	0.43	0.54	

The magnitude of the Exact Match indices presented in Table 2 vary between a low of 32% (RT, MN, for Devising Strategies) and a high of 73% (RT, CB for *Reasoning and argument*) with about half being at or in excess of 50%. This is double the rate of matches expected if ratings were random, and if each possible combination were equally likely. Nevertheless, the correlations are moderate, and the data in Table 2 show that there is quite some variation among raters, and the variability differs among rating categories. The correlations range from a low of 0.10 to a high of 0.75, with just under half being at or over 0.50. However, we do see more consistency (higher correlations) for some categories than for others. There would appear to be more consistency in the ratings applied for the *Symbols and formalism* category,

In summary values reported in Table 2 do not support reliance on the competency ratings made by any one individual. If the scheme is to support such an outcome then further work still needs to be done on the content – possibly on both the definition of the categories and the description of the rating levels.

Constructing Measures of Item Competency Demands

An alternative approach to the treatment of the ratings from the five raters is to consider them as multiple indicators of underlying latent item characteristics. Just as the responses a student makes to a set of items can be used to characterise the ability of a student, and the responses to an item made by a set of students can be used to characterise the difficulty of an item, we can consider using the set of ratings of a task's competency demands to characterise the task. That is, for each

of the six competencies we can use the five ratings as multiple indicators and test whether they hang together to form a reliable measure.

Using the nomenclature of testing, what we are doing here is measuring the competency demands of the tasks with respect to six scales (the competency demands). For each of the six scales we have five *items*, which are the ratings made for each of the tasks by the five raters. Treating each of the raters as *items* our goal is to examine whether these five items hang together well enough to support aggregation to form measures for each of the six competency demands for the tasks.

Table 3: Summary statistics for classical test theory and Rasch analyses

	Communication (R=0.80)[1]			***Devising strategies (R=0.79)***		
	Corr with Total[2]	**Corr with Rest**[3]	**MNSQ**[4]	**Corr with Total**	**Corr with Rest**	**MNSQ**
CB	**0.82**	**0.68**	**0.89**	**0.39**	**0.63**	**0.91**
KS	**0.54**	**0.36**	**1.24**	**0.81**	**0.54**	**1.06**
MN	**0.81**	**0.66**	**0.93**	**0.77**	**0.62**	**0.90**
RT	**0.81**	**0.66**	**0.89**	**0.64**	**0.59**	**1.06**
WB	**0.68**	**0.53**	**1.02**	**0.66**	**0.48**	**1.08**
	Mathematising (R=0.77)			***Representation (R=0.82)***		
	Corr with Total	**Corr with Rest**	**MNSQ**	**Corr with Total**	**Corr with Rest**	**MNSQ**
CB	**0.79**	**0.62**	**0.96**	**0.75**	**0.58**	**1.11**
KS	**0.58**	**0.35**	**1.28**	**0.75**	**0.57**	**1.18**
MN	**0.74**	**0.58**	**0.93**	**0.78**	**0.66**	**0.94**
RT	**0.87**	**0.74**	**0.80**	**0.79**	**0.64**	**1.01**
WB	**0.59**	**0.42**	**1.10**	**0.78**	**0.67**	**0.86**
	Symbols and formalism (R=0.87)			***Reasoning and argument (R=0.83)***		
	Corr with Total	**Corr with Rest**	**MNSQ**	**Corr with Total**	**Corr with Rest**	**MNSQ**
CB	**0.81**	**0.69**	**1.03**	**0.82**	**0.70**	**0.90**
KS	**0.83**	**0.71**	**1.01**	**0.74**	**0.54**	**1.31**
MN	**0.79**	**0.66**	**1.10**	**0.75**	**0.63**	**1.01**
RT	**0.84**	**0.74**	**0.89**	**0.85**	**0.73**	**0.84**
WB	**0.81**	**0.71**	**0.98**	**0.72**	**0.60**	**1.00**

Notes:

1. R is Cronbach's alpha.
2. Corr with Total is the product moment correlation between the rater's competency rating and the average rating of all raters.
3. Corr with Rest is the product moment correlation between the rater's competency rating and the average rating of the four other raters.
4. MNSQ is the weighted infit mean square fit statistic for the fit of the rater to a rasch partial credit model.

Table 3 provides some summary statistics from both classical and Rasch-based item response modelling analyses of the six possible competency scales. The classical statistics show that as scales the sets of ratings have respectable reliabilities and item-total correlations are consistently high. Similarly, with just a couple of exceptions for rater KS, the MNSQ statistics shows reasonable fit to a Rasch item response model. The typically poorer fit of rater KS is reflected in typically lower item-total and item-rest correlations for rater KS.

The worst fitting case is illustrated in Figure 2. The smooth curve shows the Rasch model's expected score – that is the mathematical expectation – for *Reasoning and argument* ratings by rater KS for tasks as a function of the overall estimated *Reasoning and argument* demand of each task. The dotted curve shows the corresponding observed reasoning and argument ratings of KS. The fact that the dotted curve is *flatter* than the model estimated curve indicates that KS tends to judge the more demanding tasks (as assessed by all raters) as a little less demanding than the other raters. Similarly rater KS tends to judge the less demanding tasks (as assessed by all raters) as a little more demanding than the other raters.

The misfit illustrated in Figure 2 is quite moderate. While clearly systematic, the difference between the observed and expected scores is not particularly large. The largest discrepancy is for the highest group where the observed average is 1.90 and the model expectation is 2.02. This would seem negligible on a scale that runs from 0 to 3.

These results provide reasonably firm support for the notion that measures of the competency demands of the PISA tasks can be constructed through a Rasch calibration of the

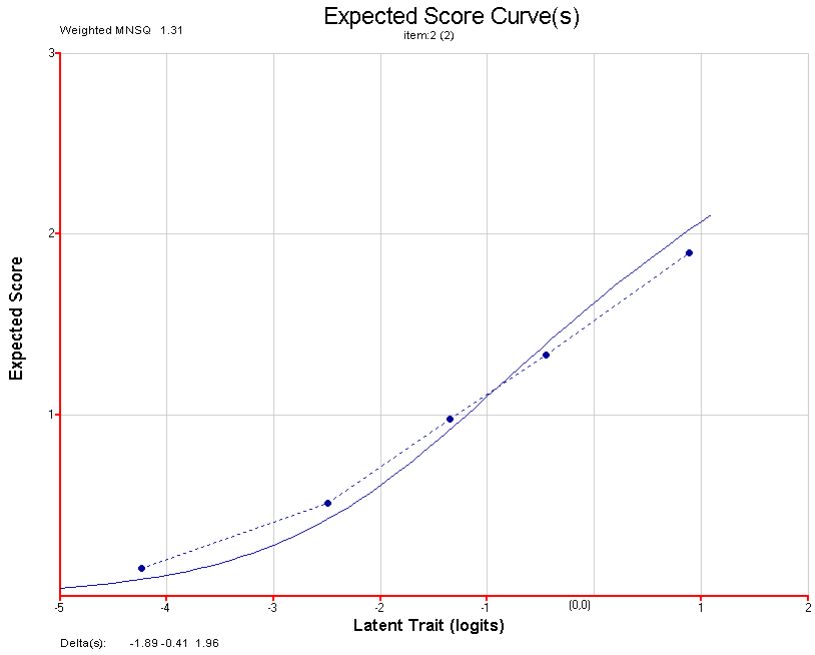


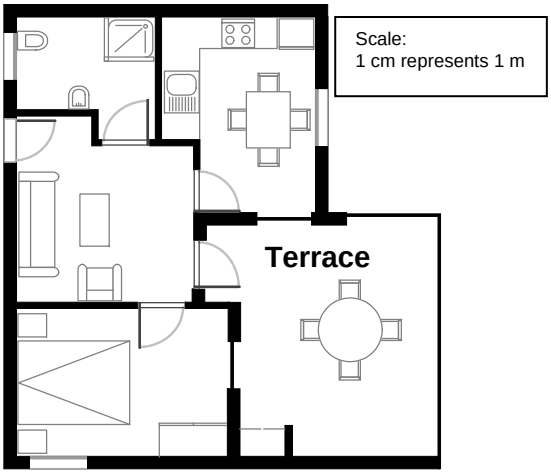
Figure 2: Comparison of Modelled and Observed Expected Score Curve – Rater KS, Competency Reasoning and Argument

Illustrating the Constructs

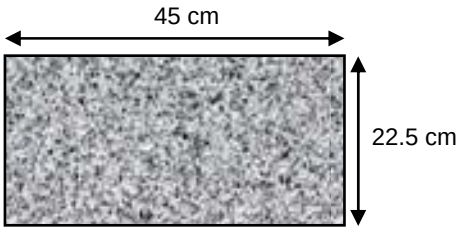
Here we provide two tasks that illustrate particular features of the rating scheme and its underlying scales. The tasks are chosen as examples that are rated relatively high on at least one of the variables, and relatively low on others.

Apartment Purchase

This is the plan of the apartment that George's parents want to purchase from a real estate agency.



The terrace's tiled floor is worn and needs to be changed. The new tiles chosen by George's parents measure 22.5 cm by 45 cm and are sold in boxes of 50 tiles each. The terrace measures about 22 m^2.



How many boxes of tiles need to be bought in order to pave the entire terrace? Show your work.

Figure 3: Sample item - Apartment Purchase

The task titled *Apartment Purchase* is rated relatively highly on the *Devising strategies* variable. The problem solver must devise a multi-stage strategy to link the dimensions of the tiles to the area of the terrace (in order to calculate the number of tiles required), and then must connect the result of that calculation to the information provided about the contents of each box. Control processes must be activated to monitor the implementation of such a sequence of strategic steps. Hence we see a relatively high level demand for *Devising strategies* (ave=2.0).

However, the *Representation* demand of this task is very low (ave=0), since little interpretation of the mathematical information presented in the text and graphics is required. The essential mathematical information is presented in the text, and no exploration or manipulation of representations is needed.

The task demands only a moderate level of activation of the *Communication* variable. Understanding what is required is reasonably straight forward, involving identifying and linking relevant elements that are nevertheless clearly presented; and the constructive communication required involves only writing a numeric result. Hence a moderate *Communication* demand (ave=1.2).

Of course some students may include in their thinking consideration of the grouting needed between tiles, and may choose to make other assumptions for example about layout of tiles, the cutting of tiles, or about parts of the floor that may not need tiles. In addition, the practical constraint related to working with whole boxes of tiles demands thinking about the relationship between the mathematical calculations and the real context. Consideration of those matters would be treated as part of the *Mathematising* demand of the task which is moderate (ave=1.4). The *Reasoning and argument* demand is relatively low (ave=1.0), involving the linking of information to make inferences, but essentially only direct reasoning within one aspect of the

problem. However, the carrying out of the area calculation, the conversion of units, and the division, involving manipulation of decimal numbers and fractions, would lead to a moderately high rating for the *Symbols and formalism* variable (ave=1.8). The sum of mean ratings for this task was 7.4.

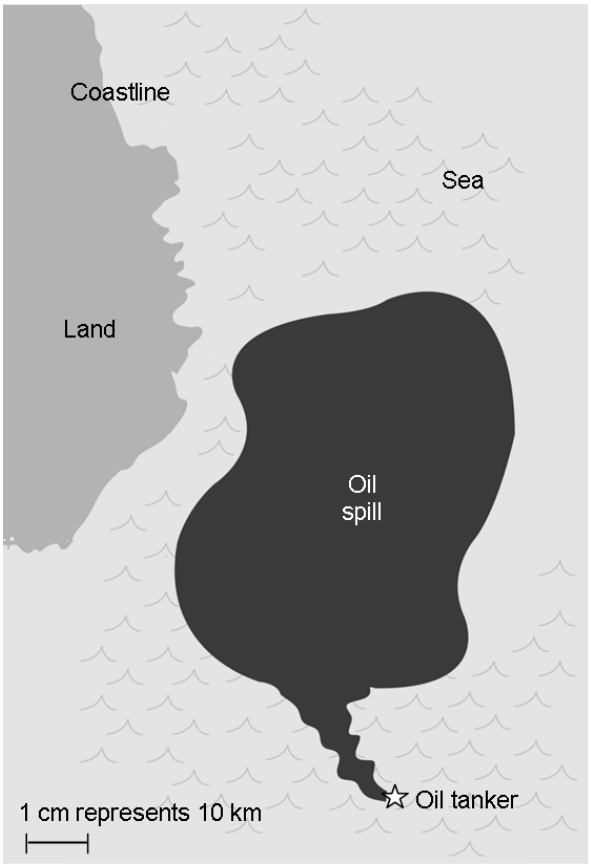
The second task to be discussed is titled *Oil Spill.* This item requires students to estimate an irregular area, and use the given scale to convert the estimate to specified units. The need to *Devise a strategy* to estimate the area is the most demanding aspect of this task (ave=2.0). Students would typically impose a geometric model for the shape, in any of a variety of possible ways, using several smaller but familiar shapes together to approximate the larger shape.

Implementing such a strategy involves transforming the real contextual information (the irregularly shaped oil spill) into mathematical form (known smaller shapes); hence there is a significant *Mathematisation* demand (ave=1.4). Using geometric knowledge to calculate the area of those known shapes, then applying the scale factor, demands activation of the *Symbols and formalism* variable (ave=1.4).

Some *Reasoning and argument* is required to optimise the approximation method, and to link the estimated area to the specified units required (ave=1.4). The diagram given in the question stimulus is a straight-forward depiction of areas, so the *Representation* demand is relatively low (ave=1.2). The information provided in the text is also straightforward and requires no interpretation or processing in order to understand what is required, and presenting a response requires only writing down a number. Hence the *Communication* demands of this task are negligible (ave=0.2). The sum of mean ratings for this item was 7.6.

Oil Spill

An oil tanker at sea struck a rock, making a hole in the oil storage tanks. The tanker was about 65 km from land. After a number of days the oil had spread, as shown on the map below.



Using the map scale, estimate the area of the oil spill in square kilometres (km^2).

Figure 4: Sample item - Oil Spill

In Figure 5 we present the mean ratings for these two tasks, to display some features of the tasks in a visual form. The average rating totals for Oil Spill were close to those for Apartment Purchase but with a rather different profile. The demand for the *Communication* variable was greater for Apartment Purchase, the *Representation* demand was higher for Oil Spill, and the two items involve similar demands for *Devising strategies*, and for *Mathematisation*. A representation of this kind

provides a visual means of identifying differences and similarities among items according to the profile of competency demand involved.

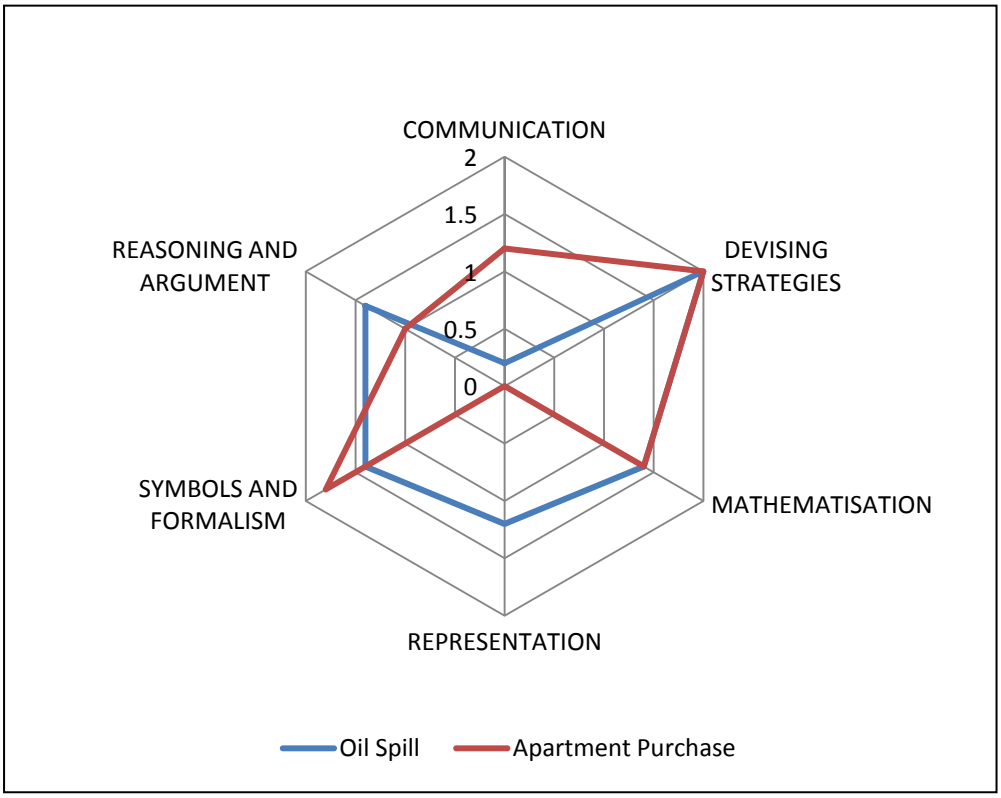


Figure 5: Graph of mean ratings of two sample items

Predicting task difficulty

Having demonstrated that the five ratings of each task's competency demands can be used to construct measures we proceed to fit a seven-dimensional item response model to examine the relationships between the six competencies and item difficulty.

Under this model it is argued that each task can be characterised by seven features – the six competency demands (Communication, Devising strategies, Mathematising, Representation, Symbols and formalism, and Reasoning and argument) and the difficulty of the task.

When the seven-dimensional model is fit the reliability estimates shown in Table 4 and the correlations between the seven dimensions in Table 5 are obtained.

The reliability values for the competency demand dimensions are a little higher than those reported in Table 3, this is because a multidimensional analysis was performed and the competencies borrow strength from each other when they are positively correlated. The reliability of the difficulty dimension is high both because of the correlations between the dimensions and because there are many student responses being used as indicators of the task difficulties.

Table 4: Estimated Correlations between Item Characteristics

	Reliability
Communication	0.764
Devising strategies	0.827
Mathematising	0.824
Representation	0.793
Symbols and formalism	0.895
Reasoning and argument	0.856
Difficulty	1.000

Table 5: Estimated Correlations between Item Characteristics

	Comm-unication	Devising Strategies	Mathe-matising	Repre-sentation	Symbols & formalism	Reasoning & arg	Difficulty
Communication	1.000	0.444	0.533	0.115	0.409	0.566	0.331
Devising Strategies		1.000	0.828	0.190	0.667	0.811	0.777
Mathematising			1.000	-0.051	0.619	0.916	0.631
Representation				1.000	0.038	0.306	0.095
Symbols and formalism					1.000	0.547	0.788
Reasoning and argument						1.000	0.669
Difficulty							1.000

Generally, these correlations are low to moderate, indicating clear discriminant validity – that is, the variables appear generally to be picking up somewhat distinct aspects of item demand. The exceptions are the correlations between *Mathematising*, *Devising strategies* and *Reasoning and argument* which are relatively high. These aspects of cognitive demand seem to be more closely related, an observation that is supported by comments from some of the raters to the effect that they tended to have trouble assigning particular aspects of the cognitive demand of some tasks separately to these variables. It could be that these variables overlap or are often activated together. Alternatively, it could be that the definitions and level descriptions are not sufficiently distinct to support a higher level of discrimination.

For these data, only one correlation is negative (between *Representation* and *Mathematising*). Perhaps for some of the tasks, when complex or multiple mathematical representations are present as part of the question, much of the mathematisation has already been done, hence high demand for one will frequently accompany low demand for the other.

Individually, four of the six variables are quite strongly related to task difficulty. The *Representation* variable, and to a lesser degree *Communication*, are not so strongly connected to task difficulty when considered alone. This is interesting, since a common observation is that one of the features of PISA task that sets them apart from many other kinds of mathematics tasks is the degree of communication required to read and understand task demands, as well as the challenging aspects of expressing arguments and conclusions in written form. This may indicate that it is not the communication demand alone that makes this variable relevant, but the way it intersects with other task demands.

Table 6: Regression Model results

Model 1 Predictors	**Coefficient**	**Std Error**	**T**
Communication	-0.170	0.002	-84.960
Devising Strategies	0.842	0.004	210.613
Mathematising	-5.591	0.020	-279.537
Representation	-0.999	0.003	-332.837
Symbols & formalism	0.489	0.001	489.370
Reasoning & argument	2.629	0.009	292.099
R-squared=0.999			

Model 2 Predictors	**Coefficient**	**Std Error**	**t**
Communication	-0.131	0.042	-3.123
Devising Strategies	0.355	0.072	4.934
Representation	-0.033	0.034	-0.972
Symbols & formalism	0.364	0.035	10.413
Reasoning & argument	0.187	0.063	2.976
R-squared=0.752			

Model 3 Predictors	**Coefficient**	**Std Error**	**t**
Devising Strategies	0.382	0.073	5.239
Symbols & formalism	0.349	0.035	9.977
Reasoning & argument	0.097	0.057	1.694
R-squared=0.739			

Model 4 Predictors	Coefficient	Std Error	t
Communication	-0.083	0.039	-2.116
Devising Strategies	0.499	0.052	9.590
Symbols & formalism	0.363	0.036	10.081
R-squared=0.741			

In Table 6, the results of various regression models are presented, regressing the competency variables as predictors on the difficulty of the tasks. Model 1 appears to generate very powerful predictors, however, the very large T values for this model are probably caused by collinearity. We conclude that this is not a good model because of high correlations between *Devising strategies*, *Mathematising* and *Reasoning and argument.*

Model 2 drops *Mathematising* because it is highly correlated with both *Devising strategies* and *Reasoning and argument* and because its large negative coefficient suggests a suppressor effect. The remaining variables predict some 75% of the variability in task difficulty.

Model 3 drops *Communication* and *Representation*, simply because they had negative coefficients in both models 1 and 2. The remaining variables predict some 74% of the variability in task difficulty.

Finally, model 4 includes a single representative, *Devising strategies*, from the strongly interrelated group *Devising strategies*, *Mathematising* and *Reasoning and argument.* It also includes *Communication* and *Symbols & Fomalism,* because they were statistically significant when used in the preceding models. *Representation* is omitted because it was not statistically significant. The remaining variables again predict some 74% of the variability in task difficulty.

For this final model the standardised coefficients for the three variables, *Communication, Devising strategies* and *Symbols & formalism,* are –0.112, 0.461 and 0.526, respectively. So, all other things equal, a one standard deviation increase in either the *Devising strategies* and *Symbols & formalism* demand of a task leads to about a half standard deviation in increase in the difficulty of the task. In the case of *Communication*, all other things equal, a one standard deviation increase leads to a decrease of a tenth of a standard deviation in task difficulty.

These results are very encouraging, suggesting that the competency demands of the PISA tasks are very powerful predictors of their difficulty.

Conclusion and discussion

This research has the potential to significantly advance our understanding of the central drivers of item difficulty in mathematics. It carries implications for understanding and describing growth in mathematical competence, and therefore for the place of the competencies in the mathematics curriculum and in the daily work of mathematics teachers. These implications are likely to be relevant far more widely than in the PISA context alone.

Several directions for developing this research are evident. The task difficulty data in this analysis are derived from Australian student responses to PISA tasks. It will be important to check whether any differences in the observed patterns emerge when data from other countries are analysed. It

would be critical to cross-validate the finding reported using data from more countries, since it is possible that the findings regarding task difficulty are to some extent curriculum dependent.

Similarly, it will be useful to extend this work to involve a wider variety of raters. It could be that those involved so far have generated the observed data partly as a function of their previous involvement with the PISA project and their familiarity with PISA items. For this work to be extended to other raters with less direct experience of the PISA context, clear explanatory and training material will need to be developed and validated.

A further extension of potential interest will be to ascertain the extent to which similar results might be observed using other tasks. To what extent are these finding generalisable across the universe of different kinds of mathematics tasks?

And of course there is evidence that the definitions of the variables and the descriptions of levels of activation of those variables are not yet well enough developed to support distinct judgments to be made. The overlap among the variables may be reduced with further attention to the definitions and the wording of the descriptions. Indeed, the first action taken by the raters following the most recent round of item rating was to meet and discuss the definitions and descriptions, and already refinements have been made to the rubrics with a view to making them clearer and more distinct.

One possible benefit of that kind of attention might be to reduce the number of raters needed to generate useful ratings. It would be ideal if the rubric could be improved to such a degree that more useful analysis could be generated by an individual rater.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). *ACER ConQuest Version 3: Generalised item response modelling software* [computer program]. Camberwell: Australian Council for Educational Research.

Adams, R. J., & Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21, 1-24*

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.